HD Webcam TrueLife Display

Week-7

Logistic Regression - Predicting the placements

2022 MARCH

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

1 FRIDAY
2022 APRIL

WEEK 14
091-274

Categorical predictions

- Placement proces B-Schools (Buisness Schools) facilitates to pick a job of their choice (amongst the available profiles).
- The student attributes (academic performance, prior experience, internships) is expected to have a strong bearing on the outcome of the placement proces.
- The outcome variable in the example is binary - a student either gets a job or she doesn't
- However the idea is more generic - the outcome variable could have several categorical values.

X_1 = Academic performance during MBA
X_2 = Industry experience prior to joining the MBA program
X_3 = Academic performance during the undergraduate degree
X_4 = Participation in the co-curricular and extra-curricular activities.

$Y = 1$ if student get placed else zero.

HD Webcam TrueLife Display

Week-7

Logistic Regression - Predicting the placements

2022 MARCH

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

1 FRIDAY
2022 APRIL

WEEK 14
091-274

Categorical predictions

- Placement proces B-Schools (Buisness Schools) facilitates to pick a job of their choice (amongst the available profiles).
- The student attributes (academic performance, prior experience, internships) is expected to have a strong bearing on the outcome of the placement proces.
- The outcome variable in the example is binary - a student either gets a job or she doesn't
- However the idea is more generic - the outcome variable could have several categorical values.

X_1 = Academic performance during MBA
X_2 = Industry experience prior to joining the MBA program
X_3 = Academic performance during the undergraduate degree
X_4 = Participation in the co-curricular and extra-curricular activities.

$Y = 1$ if student get placed else zero.

Predicting the placements

- However, we need to pay attention to our response variable.
- Since the response variable is binary (or generically speaking, categorical), we can't use the regular expression method and expression.

Solution method: Regression

- If was modeled as a multiple linear regression, we would have

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

- Since, our Y is binary, assumptions of the regression model won't hold and we won't get good predictions
- That is: $Pr\{Y=1\}$ as a predictor. Then, our response variable has values between 0 and 1.
- However, if we calculate ODDs, then we can get out of these limits.

$$\text{Odds (success in placements)} = \frac{Pr(Y=1)}{Pr(Y=0)}.$$

- More commonly, log values are used. That is, log of the odds.
- As a result, we have (dropping error terms)

$$\text{Log(Odds)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

↳ find

Taking antilog

$$Odds = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$$

$$P_r(Y=1) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}$$

$P_r(Y=1)$ → forecasted probability

- Now we can run the regression model and estimate the regression coefficients (the β's)
- The objective function used for this estimation: maximization of the log-likelihood. That is the log of probability of the correct prediction.

Regression: Calculations explained

- Notice that this regression will give us the "forecasted" probability that the student will be placed ($Pr\{Y=1\}$).
- However we want the value for our response variable (Y). And not the probability that Y will take on certain value.
- Towards that, we define a threshold probability - if the forecasted probability value is above the threshold, we say that $Y=1$.
- On the other hand, if the forecasted probability is below the threshold, we can say that $Y=0$ (the student won't be placed).

MAY 2022

M	T	W	T	F	S	S
30	31					1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

TUESDAY
APRIL 2022
5

WEEK 15
095-270

Evaluation of the Model

How we judge whether this logistic regression model is a good?

- Typical statistical indicators: (generally based on the log-likelihood) – deviance, R^2, and information criteria (Akaike and Baye's).
- Some of them have a threshold (often χ^2 based statistic) or sometime it is higher-the better type (e.g. R^2)

- Other performance indicators: (generally based on correct identification) – Accuracy, Precision, Recall
- Obviously, they are all large-the-better type performance indicators.

- Accuracy :- Measure of the total number of predictions a model gets right, including both True +ves and True -ves.

- Recall :- Indicates the percentage of the response values (that we are interested in) were actually captured by the model.

- Precision :- Measures the percentage of the predicted response values (that we are interested in) that were correct.

6 WEDNESDAY
2022 APRIL

2022 MARCH

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

WEEK 15
096-269

The performance measures can be interpreted as:

- **Accuracy**: The ratio of the number of times predicted and actual Y values matched (for both Y=0 and Y=1) to the total observations in the sample.
- **Recall**: The ratio of number of times the prediction for Y was 1, to the total number of instances in the sample where Y was actually 1.
- **Precision**: The ratio of the number of times the actual Y was 1, to the total number of instances where the prediction for Y was 1